DOCUMENT RESUME

ED 159 176 TM 007 327

AUTHOR Fishbein, Ronald L.
TITLE The Use of Non-Normed Tests in the ESEA Title I Evaluation and Reporting System: Some Technical and Policy Issues.
PUB DATE Mar 78
NOTE 15p.; Paper presented at the Annual Meeting of the American Educational Research Association (62nd, Toronto, Ontario, Canada, March 27-31, 1978)
AVAILABLE FROM Ronald L. Fishbein, U.S. Office of Education, FOB-6-Room 3044, 400 Maryland Avenue S.W., Washington, D.C. 20202 (no charge)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Achievement Gains; Compensatory Education Programs; Criterion Referenced Tests; Educational Objectives; Elementary Secondary Education; Equated Scores; *Mathematical Models; *National Norms; *Norm Referenced Tests; Policy; *Program Evaluation; *Scoring; *Standardized Tests; Test Interpretation
IDENTIFIERS *Elementary Secondary Education Act Title I

ABSTRACT

Intended primarily for local education agency evaluators and state education policy makers, this paper explores several technical and policy issues related to the use of non-normed tests in the Elementary and Secondary Education Act Title I Evaluation and Reporting System (TIERS). The technical issues discussed were: (1) the assumption, necessary to implement the control group (B2) and regression (C2) models, that the ratio between the sample and population standard deviations on the nationally normed test and the test without national norms is the same; (2) the linear transformation used in models B2 and C2 to convert the local treatment effect to normal curve equivalents, along with possible alternative procedures for using non-normed tests in the Title I system; and (3) floor and ceiling effects in relation to model A2 evaluation (the norm-referenced model which can be implemented with a test without national norms through an abbreviated version of equipercentile equating). The policy issues discussed were the scope of the curriculum which can be tested when using a model A2, B2 or C2 evaluation; and additional procedures that might be undertaken. It was concluded that, in the vast majority of cases, only a limited set of program objectives can be evaluated with a standard model A2, B2 or C2 implementation. It was recommended that those employing a non-normed version of the Title I evaluation system consider incorporating a normative assessment component. (Author/CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED159176

The Use of Non-Normed Tests in the ESEA Title I Evaluation and Reporting System Some Technical and Policy Issues

Ronald L. Fishbein
United States Office
of Education
Office of Planning, Budgeting
and Evaluation

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Ronald L. Fishbein

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

TM007 327

Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978

Session 14.15

The use of non-normed tests in the Title I evaluation and reporting system (TIERS) was first presented to the educational research community at the 1976, San Francisco, AERA. At that meeting I presented a paper, coauthored by Janice Anderson, outlining procedures for using tests without national norms in the control group (B_2) and special regression (C_2) evaluation models. Since that time, these procedures have been documented in materials written by staff members from RMC Research Corporation (Mountain View, California) and in materials developed by the ESEA Title I Regional Technical Assistance Centers (TACs). RMC and the TACs have also developed materials about model A_2, the norm-referenced model which can be implemented with a test without national norms through an abbreviated version of equipercentile equating (Bessey, Rosen, Chiang, & Tallmadge, 1976).

There are several technical and policy issues which could be discussed in detail concerning the use of non-normed tests in the TIERS. However, in this paper I attempt to limit the discussion to an overview of issues particularly pertinent to local education agency (LEA) evaluators and SEA policy-makers. The reader is cautioned that the issues have been simplified for discussion purposes. For those of you who will be attending the National Assessment of Educational Progress (NAEP) conference in June, there will be a symposium on the use of non-normed tests in the TIERS, with several participants focusing on the technical issues.

Technical Issues

Models B_2 and C_2

Equality of the ratios of nrt and crt population/sample variances.

Models B_2 and C_2 require that the treatment effect calculated on the criterion-referenced test (CRT)[1] be converted to a normal curve equivalent (NCE) gain. This may be done by using Equation 1, where 21.06 is the standard

$$TEnce = 21.06 \left(\frac{\bar{Y} - \hat{\bar{Y}}}{\sigma_{crt}}\right) \quad (1)$$

deviation of the NCE scale, $\bar{Y}$ is the treatment group's observed posttest mean, $\hat{\bar{Y}}$ is the treatment group's expected mean posttest score, and σ_{crt} is the standard deviation of the appropriate national population on the CRT.

One can estimate σ_{crt} by assuming that the ratio between the sample and population standard deviations on the nationally normed test (NRT) is the same as the ratio between the sample and population standard deviations on the CRT. That is, one assumes the validity of Equation 2 and solves

$$\frac{\hat{\sigma}_{crt}}{S_{crt}} = \frac{\sigma_{nrt}}{S_{nrt}} \quad (2)$$

for $\hat{\sigma}_{crt}$, where

$\hat{\sigma}_{crt}$ = the estimated standard deviation of the appropriate national population on the CRT,

S_{crt} = the standard deviation of the treatment group's posttest scores on the CRT,

σ_{nrt} = the standard deviation of the appropriate national population on the norm-referenced test (to implement either model B_2 or C_2 a test with national norms must be administered to at least a representative sample of the treatment group to obtain S_{nrt}; σ_{nrt} is found in the publisher's manual),

S_{nrt} = the standard deviation of the treatment group's scores on the norm-referenced test.

The value $\hat{\sigma}_{crt}$ may be inserted into Equation 1, and the treatment effect in NCEs calculated.

As Bessey et al. (1976) have indicated, the assumption made in Equation 2 can be tested empirically by collecting scores on the CRT from a representative national sample and comparing the standard deviation of those scores with the standard deviation that would have been predicted ($\hat{\sigma}_{crt}$) from Equation 2. In such a study, one would have to assume the equality of the NRT and CRT national samples. Although not strictly a test of the assumption made in Equation 2, one could administer two NRTs in a district and predict one of the national standard deviations, comparing the predicted value with the value provided in the publisher's manual.[2] To my knowledge, no study has been undertaken testing the assumption made in Equation 2.[3]

Linear transformation. Models B_2 and C_2 use a linear transformation to translate the gain on the CRT to NCEs. As pointed out by Jaeger (1977), the "normalized z-score method and linear transformation method of computing NCE scores will provide equivalent results only when the norm distribution of a test is normally distributed (p. 59)." That is, a linear transformation that does not normalize scores maintains the shape of the original distribution.

An important question for the Title I evaluator is: Will the use of the linear procedure of computing the NCE Treatment Effect in models B_2 and C_2 give different enough results from models B_1 and C_1 for the evaluator to be concerned? Stated differently, if the CRTs used with a B_2 or C_2 evaluation had national norms, and a B_1 or C_1 evaluation were conducted, would the results be "equivalent?" This issue could be examined by administering two tests with national norms (one or both could be a CRT with national norms), using one as the CRT in a model B_2 or C_2 evaluation. The results could be compared with results obtained in a model B_1 or C_1 evaluation using the nationally normed test previously used as the CRT. To my knowledge, such a study has not been undertaken.

G. Kasten Tallmadge of RMC has made several relevant points concerning the linear transformation issue. In a July, 1977 letter to John Pyecha of Research Triangle Institute, Tallmadge indicated that the gains (treatment effects) in models B_2 and C_2 are always treatment-group posttest scores minus control (comparison) posttest or predicted posttest scores. Furthermore, Tallmadge noted that if the scores on the CRT are skewed, for example,

both the treatment and control (comparison) group distributions should be skewed in the same direction. Although there would probably be some differences, again, the question to consider is whether the differences would be large enough to affect substantially the NCE gain.

Test equating. Evaluators employing model B_2 or C_2 who are concerned about some of the potential hazards of transforming the CRT treatment effect to NCEs via a linear transformation might instead consider equating the NRT and CRT(s). One would then have, in effect, a model B_1 or C_1 .

Before proceeding with such a task by one of the commonly used equating procedures, such as those described by Angoff (1971), one should consider very carefully the effects of violating the assumptions of the equating procedure chosen. Angoff alerts us that "equating, or the derivation of equivalent scores, concerns itself with the problem of unique conversions which may be derived only across tests which are parallel (p.563)." Parallel forms, of course, should have correlations approaching 1.0. Although a professionally developed CRT would most likely be substantially correlated with a NRT in the same content area, they will not be parallel or nominally parallel (as defined by Lord & Novick, 1968). In this writer's experience, such tests will usually have maximum correlations of. 70.

With the development of latent-trait equating procedures (the Rasch model, for example), some of the pitfalls of equating NRTs and CRTs with traditional procedures may be avoided. But the reader is cautioned that the implementation of such a procedure in a practical situation is still experimental.

Another related possibility would be to develop a regression equation to predict NRT scores from the CRT(s) (or CRT(s) scores plus other variables). The predicted scores could then be used in a model A_1, B_1 or C_1. However, unless the tests are highly correlated, such a procedure would be very tenuous.

Model A_2

According to Bessey et al. (1976), model A_2 requires that a nationally normed test be administered at pretest time. Then, an abbreviated version of equipercentile equating is "used to 'bridge the gap' between the observed median posttest score on the non-normed tests, its normed test counterpart, and the corresponding national norm percentile value (p.18)."

Tallmadge (1976), has indicated that when using CRTs with the TIERS, there must be enough easy items on the pretest to provide a reliable indication of how much the pupils know, and enough difficult items on the posttest to determine how much the pupils know at that time -- in other words, one should avoid ceiling and floor effects. For the

purposes of this paper, it is sufficient to indicate that a ceiling or floor on any of the three instruments used in a model A_2 evaluation may be a potential serious source of error to the evaluation. As just one example, a large ceiling effect on the CRT posttest and a large floor effect on the CRT pretest may make it impossible to use the model -- there may be no cumulative frequency on the pretest CRT that corresponds to the median raw score value on the CRT posttest. In the author's opinion, one would be wise to heed Tallmadge's advice about choosing tests with sufficient score variance.

Policy Issues

Those of us who have been involved in the operation of a large-scale CRT testing program are aware that persons encouraging the use of such tests point out that when an objectives - based instrument is developed, it may give a more accurate indication of how well the students have mastered the program objectives than a commercially produced NRT. If this is true, then, from the perspective of limiting the purposes of the program to improvement in the precise instructional objectives taught, CRTs may be more appropriate for evaluating instructional programs than commercially produced global NRTs.

For the purpose of evaluating an instructional program in terms of how well the students have mastered the program objectives, the instruments used, whether CRTs or NRTs, should match the curriculum objectives as closely as possible. However, the evaluator and program designer of a

compensatory educational program who have chosen to use CRTs, perhaps for the precise reasons just outlined, might wish to consider some of the issues which follow.

Scope of Curriculum and CRTs

If the objectives of a program are narrowly defined, one could develop instruments to measure mastery of those objectives and perform a technically correct evaluation using model A_2, B_2 or C_2. However, one should ponder the meaning of the NCE gain (and/or the treatment effect in the local metric) which has been demonstrated. If the gain is large, this means that the students have shown more growth on the narrow set of objectives than would have been expected without the special program. But what about the possibility that because so much instructional time has been devoted to meeting the narrow goals of the program, other important curriculum areas have been neglected? In addressing this issue, Tallmadge and Horst (1977), gave the example of a compensatory educational reading program that concentrated on phonics. Such a program might show a large NCE gain, but the students may be no better "readers," maybe even worse (better being defined as achievement above what would be expected from regular schooling).

The reader may be thinking that some curricula incorporate many objectives and do not have the problem of neglecting important areas. Furthermore, the reader may correctly conclude that a compensatory educational program geared to a broadly based set of objectives can be evaluated with CRTs. Matrix sampling procedures (Shoemaker, 1973) will permit the evaluator to limit the time required for testing and, at the same

time, measure performance on many objectives. Unfortunately, for the vast majority of the almost 14,000 school districts receiving Title I funding, it is not feasible to implement matrix sampling procedures. The local evaluator will usually not have the technical expertise to implement the procedure, nor the appropriate tests. The technical problems can be overcome with outside assistance, but somebody must pay for the development of the appropriate tests. Unless the state education agency is developing item banks for state-level objectives, or unless the district is rich enough to pay for the development of appropriate tests, the district will be forced to limit severely the number of objectives tested in a CRT used for Title I evaluation. This will be the situation for the vast majority of Title I districts considering using a model A_2, B_2 or C_2 evaluation design, regardless of the scope of the curriculum.

Policy and Program Objectives

At this point the reader may be looking for guidance. Perhaps it would be helpful to consider the distinction between Title I program objectives and Title I policy objectives. Program objectives can be defined as the specific instructional objectives of the program -- the instructional tasks that the student will hopefully master at the end of the program. Policy objectives, such as "improvement in reading" are more global and may be met because of successful instructional programs. A policy objective is more difficult to achieve than a program objective. For the vast majority of districts that consider using CRTs for Title I evaluation, only growth in program objectives will be measured unless the measurement of policy objectives is also incorporated into the evaluation.

From discussions with persons in policy-making positions at the local, state and federal levels the author has found almost unanimous agreement that the major goal of Title I is overall relative improvement in reading and mathematics. That is, if the purpose of Title I is to compensate for a disadvantaged background, then compensation is demonstrated when relative standing has improved.

If one is willing to concede that policy goals may differ from program goals and that policy goals should eventually be reached as a result of successful programs, then it would seem logical to measure the achievement of both types of goals. It is not necessary to do this yearly, but, in the opinion of this writer, the evaluator using a model A_2, B_2 or C_2 design should, at some time, attempt to determine whether changes in relative standing (particularly for students who have been in a district's Title I program for several years) have occurred. One easy way to do this is to determine the percentile or NCE of the mean score on the NRT used in the evaluation (the NRT must be given anyway and one should attempt to give it at an empirical norm point) for the Title I group in each grade and some subgroups within each grade's Title I population. Subgroups might be formed from students who have been in the district's program for given numbers of years. The suggestion being made is that it is possible, with little additional effort, to build into the Title I evaluation an assessment of relative standing (for example, a fall-spring model A_2, B_2 or C_2, and a fall - fall model A_1) If enough students were available, one could conduct a longitudinal model A_1 evaluation.

Summary

In this paper several technical and policy issues related to the use of non-normed tests in the ESEA Title I Evaluation and Reporting System were explored.

The technical issues discussed were (1) the assumption, necessary to implement models B_2 and C_2, that the ratio between the sample and population standard deviations on the nationally normed test and the test without national norms is the same (assuming the same population and sample for each test); (2) the linear transformation used in models B_2 and C_2 to convert the local treatment effect to NCEs, along with possible alternative procedures for using non-normed tests in the Title I system; and (3) floor and ceiling effects in relation to model A_2 evaluations.

The policy issues discussed were the scope of the curriculum possible to test when using a model A_2, B_2 or C_2 evaluation, and additional procedures that might be undertaken. It was concluded, that in the vast majority of cases, only a limited set of program objectives can be evaluated with a standard model A_2, B_2 or C_2 implementation. It was recommended that those employing a non-normed version of the Title I evaluation system consider incorporating a normative assessment component.

Footnotes

[1]CRT will be used in place of non-normed in the remainder of the paper. Although a CRT can be normed, most of the tests without national norms used in Title I evaluations are CRTS. Also, it will be assumed that all norm-referenced tests (NRTs) have national norms.

[2]A truer test of Equation 2 would be to follow this same procedure with a CRT with already existing national norms.

[3]Since this paper was written I have learned that Long, J., Horwitz. S., & De Vito, P., will present a paper entitled, An empirical investigation of the ESEA Title I evaluation systems procedures for use with criterion referenced tests, in session 32.11. Long et al. empirically tested Equation 2 using the California Achievement Test as the CRT and the California Test of Basic Skills as the NRT.

[4]Title I funds can be used to pay for the purchase of tests for Title I children (not for control-comparison group students or for the development of tests). But at least for the next few years, it seems improbable that a district will be able to buy items that match its specific objectives.

References

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1971.

Bessey, B. L., Rosen, L. D., Chiang, A., & Tallmadge, G. K. Further documentation of State ESEA Title I reporting models and their technical assistance requirements: Phase I report. Mountain View, California: RMC Research Corporation, 1976.

Jaeger, R. M. On combining achievement test data through NCE scaled scores. Report prepared for the Research Traingle Institute Center for Educational Research and Evaluation, Research Triangle Park, North Carolina, October, 1977.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading Mass.: Addison-Wesley Publishing Co., 1968.

Shoemaker, D. M. Principals and procedures of multiple matrix sampling. Cambridge, Mass.: Ballinger Publishing Co., 1973.

Tallmadge, G. K., & Horst, D. P. Using the data from state and local ESEA Title I reports. Paper presented at the annual meeting of the American Educational Research Association, New York, New York, April 1977.

Tallmadge, G. K. Criterion-referenced tests: ESEA Title I evaluation and reporting system. Mountain View, CA.: RMC Research Corporation, October, 1976. (RMC Technical Paper No. 11)